

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-04 15:10:16

PAGE 1

REFERENCE NO: 216

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Author Names & Affiliations

- Dmitry Mozzherin - Illinois Natural History Survey

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Biodiversity Informatics,

Title of Submission

A persistent Names-based Cyberinfrastructure for Biodiversity Sciences

Abstract (maximum ~200 words).

Biologists are under pressure to address large scale questions, such as the origins and spread of pathogens, parasites or invasive species; or how distributions of large number of species might change under certain projected conditions. These often have an historical context. The usual approaches to research are not suited to these questions. A cyberinfrastructure that can transmit digital content, and across which reasoning can occur, offers opportunities for new kinds of research suited to large scale issues. Progress with digital biodiversity projects such as Open Tree of Life, GlobalNames, iDigBio and ZooBank has set the stage for a major step towards an infrastructure. An implementable component will exploit names and other labels of organisms as metadata along with unifying comprehensive phylogenies as ontologies. Such a structure will be able to discover content - whether in molecular databases or in digital versions of documents 2 centuries old, and intelligently combine it for re-use across a very large subset of all biodiversity sciences. Such a structure will provide new roles for taxonomists and phylogeneticists, and usher in a new style of biodiversity research.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Vision:

The National Academies envision a 'New Biology' (1) reshaped by the digitization of the discipline. The International Union of Biological Sciences argues we are not well prepared for larger scale challenges and we must invest in unifying systems (2). We can address this challenge with an appropriate cyberinfrastructure. Our vision extends the agendas of digital biodiversity projects to create a names-based component (NBI) of the future Cyberinfrastructure for Biodiversity Sciences. NBI intelligently manages scientific names and other identifiers of taxa as metadata so that associated data can be discovered and organized. A unified comprehensive phylogenetic framework acts as a navigable framework to make information available for re-use in a discipline-centred context.

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-04 15:10:16

PAGE 2

REFERENCE NO: 216

Pre-existing elements of NBI have emerged in response to research needs, but the role of NBI is not to solve a particular research problem. Rather, like GenBank, GPS, particle colliders, or radio telescopes, NBI will enable new waves of investigative techniques, create new research agendas, and give new roles to phylogeneticists, taxonomists, and those working with literature, museum and herbarium specimens.

Research challenge

The scope and scale of biology presents a daunting challenge for the New (digital) Biology. Biological processes range over 35 orders of magnitude of time (from less than a femtosecond to over 3 billion years) and involve particles which extend over 23 orders of magnitude of size from subatomic elements to the complete biosphere. These two axes define “Nature’s envelope” that embraces all biology. It is a grid with over 800 sectors each embracing events across ten-fold ranges of size and time. Within this, about a third of the sectors are actually occupied by biological processes.

What we know of events in the envelope is fragmented across thousands of sub-disciplines (2). Insights are made available in over 5000 journals and 800,000 articles per year (3). In areas such as organismic biology, ecology, and evolution, the identities of the organisms and hence their names are important. In these areas, historical information is in over 500 million printed pages being digitized through BHL and its global partners. Over 700 million occurrence records from over 30,000 data sets are compiled by GBIF, Map of Life, OBIS, etc., and an estimated 3 billion specimens are curated in over 55,000 museums and 3500 herbaria in over 200 countries. With at least 2.5 million described extant and extinct species, much information is in a long tail of hard-to-access small data sources (4). The almost universal use of scientific names in these sources for over 250 years allows NBI to use names as metadata to discover content (5).

Traditional approaches are being eclipsed by molecular technologies, with almost 250 million sequences in GenBank. While the future clearly lies with genetic identification, management of names remain essential. Names give us access to historical records, they are required in taxonomic practices and have an irreplaceable value in communication. Only about 40% of GenBank content is labelled with scientific names (6). Thanks to integrative ‘Big Trees’ created through OTOL (7), different kinds of labels for taxa - such as molecular identifiers - can be managed together with names, as can phylogenies and classifications (6). This allows NBI to take advantage of the precision and convenience of molecular techniques.

The use of scientific names is not without problems. Mis-spellings, variations in style, synonymies, concept changes, increasing relevance of molecular identifiers, and misidentifications make the use of names cumbersome. New tools and strategies are emerging to ease the process (5, 6). ‘Reconciliation’ addresses the ‘many names for one taxon’ problem (5). It interconnects alternative name-strings for the same taxon such that a query initiated with one name is automatically expanded to retrieve data under all names. Resolution (8) enhances this by presenting information about the taxon using the name that comes from a preferred authoritative source.

Nature of the future cyberinfrastructure

Without automated indexing, researchers must invest considerable effort to find and acquire information from the myriad of sources. Genbank provides a model of how this can be overcome as GenBank has become a one-stop-shop for molecular information. Around GenBank has emerged an ecosystem of more specialist environments and other systems to better meet the needs of users.

We generalize the GenBank solution as modules. Each takes responsibility for managing information in an area of expertise. At the core is a node that discovers and acquires data, assigns provenance, and makes content available in standard formats. To serve the larger community, nodes must acquire information from across the globe. Nodes can make content available to users via web interfaces, APIs, and via the Linked Open Data cloud. Cloud use by the European Open Biodiversity Knowledge Management System includes about 3 million triples, and within the US, the Data Conservancy is exploring use of RMAP funded through the Sloan Foundation. The use of custom filters (location, date, taxon etc), or aggregation (using phylogenies to expand a request from, for example, *Drosophila* to all fruit flies) will deliver custom packages and enhance user satisfaction. By adding universal identifiers (UUIDs) to data items, nodes can engage tools such as Kurator and FilteredPush (9, 10) as effective mechanisms by which users address the completeness, accuracy or other issues about data quality. UUIDs can track the movement and re-use of data, and offer a mechanism to provide credit to all contributors (6, 11).

Adjacent nodes with overlapping or complementary content will cooperate using agreed transfer formats, ontologies and metadata. The interactions that define the landscape will be driven by research agendas that require particular combinations of data. Components for the example of NBI are provided in Part 2 of this document. As new needs and new opportunities arise, the modular composition coupled with standardized ontologies will allow the network to evolve and adapt.

We believe that the stage is now set for implementing a names-based cyberinfrastructure that will enhance visibility of, access to, and re-use of information from half of the envelope of biological activities. This development, will provide new agendas, new purpose and greater relevance for Global Names, OTOL, iDigBio, GenBank, annotation systems such as FilteredPush and for taxonomists.

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-04 15:10:16

PAGE 3

REFERENCE NO: 216

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Elements needed for a production NBI follow. Modules may arise from projects such as BHL, iDigBio, Global Names, Map of Life, OTOL. To serve as infrastructure, nodes need to perform at a 'Production' level of service (below). Existing projects do not offer appropriate levels of service, lack much available content, cannot guarantee longevity, and lack open-ended interoperability (12). An exception to the interoperability criterion is the Open Biodiversity Knowledge System (OBKMS) from the pro-iBiosphere project in Europe.

Levels of Service

- (1) Proof of concept (PoC). Demonstrates that a concept can be implemented. Low cost, typically put together by an individual, often within the context of a research grant. Accessible to insiders. Scope is limited
- (2) Prototype. Openly accessible service, usually with more content, dedicated hardware, robust software, and capable of satisfying up to 80% of user requests. Almost always created within a research context, but without long term finances. Most familiar services are at this level.
- (3) Production. Openly accessible service. Aims to satisfy at least 80% of user requests. Acts as infrastructure, but persistence is not guaranteed.
- (4) Professional. Users satisfied with 95% or more responses. Major investment need to ensure stability, speed, best algorithms, access pathways, and appropriate content. Also incurs maintenance costs.

Sources of names

NBI needs to gather all name-strings (the characters and spaces that make up a label) used to refer to any taxon from the following.

1. Literature. The primary source. Digital repositories (from which a literature sub-node may emerge) include Biodiversity Heritage Library, BioStor, JSTOR, Biodiversity Literature Repository, BIOSIS, or the US National Library of Medicine.
2. Specimens in museums, herbaria and culture collections contain names that are critical to taxonomy, the correct identification of taxa, and include valuable historical occurrence data. Digital specimen environments such as iDigBio could act as a sub-node. Risks to existing collections make this urgent (14).
3. On line databases. There are likely thousands that cover the full extent of Nature's envelope. High-use databases that use identifiers other than names, and sources of vernacular names need high priority. Absent elements include comprehensive lists of synonyms and georeferenced vernacular names.
4. Taxonomies and phylogenies These form biologically meaningful compilations of names and can be combined as supertrees as demonstrated by OTOL (8). Comprehensive supertrees can be exploited to identify and disambiguate homonyms. Logical formalisms can reason across trees to identify potential synonyms, an area of content that is yet to be effectively compiled. Trees enable navigation around content, allow aggregative queries (show me all information on any flat mite)

Internal Databases

1. A repository for all name-strings. GNI (<http://gni.globalnames.org/>) is ready and currently has over 20 million name-strings.
2. A registry for information about sources, taxonomies and phylogenies. This will give access to metadata for attribution, metrics about re-use and allow data quality feedback, synonymy information etc. GBIF Classification Bank is an example.

Node tools

1. Name recognition and discovery tools identify names in sources (e.g. GNRD <https://github.com/GlobalNamesArchitecture/gnrd>).
2. Names normalization. Requires parsing of name strings, now possible with Global Names parser (<https://github.com/GlobalNamesArchitecture/gnparser-paper>) and use selected elements to create preferred versions of names
3. Identifier minting service. All semantically separable data elements require UUIDs. UUID5 services allow other players to mint the same

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-04 15:10:16

PAGE 4

REFERENCE NO: 216

UUID for the same name-string.

4. Reconciliation services, to bring together all identifiers used for a taxon. The process involves cross mapping algorithms and expert input.
5. Resolution services / authority control allow data objects to be presented under a name endorsed by a source preferred by the client. Proof of concept created within iPlant (8).
6. Annotation services With browser plugins that allow users to add comments to and track movement of any data element. This system captures feedback on data quality and collects re-use metrics
7. Export services. Web interface, APIs, and release of triples for open data cloud use. All including accepted export protocols and standard data formats as well as user-friendly formats.
8. Taxonomically Intelligent Alert Services. Increase engagement by reporting new data elements or annotations for particular names or clade to registrants.
9. Cross-linking services. Devices to recognize names in digital sources, and automatically link to other services. Proof of concept, NameSpotter (<https://chrome.google.com/webstore/detail/namespacepotter/pogmoobpbggadhlleijfpjgnpkjdnhn>).
10. User Interface for Editing. This is needed to help address problems such as name changes as a result of taxonomic and phylogenetic insights, or because sources are incomplete or contain errors, or the paucity of on-line adequate synonymy information. An interface allows users to introduce new content, make improvements and corrections.
11. Upload of content. Accept lists and hierarchies in many formats, convert to standard formats, automatically populate the internal databases, reconciliation groups, and big trees; alert users to availability of new content.
12. Capacity to add synonyms, vernacular names, other identifiers and surrogates.
13. Open reviewing, annotation and correction of content - tools to improve reconciliation groups, homonym disambiguation, name-string errors, etc. using controlled vocabularies.
14. Display of comments. The interface to show new comments
15. Topology management. Edit the topology of the parent-child structure to create custom trees

Proof of concept / prototypes such as GNITE, Symbiota, TaxonWorks, Scratchpads, etc; can be used to establish requirements

Nomenclator subnode

A nomenclator compiles code compliant scientific names, the author(s) of names, the date of authorship, reference to the publication, reference to type material, and nature of nomenclatural acts. Nomenclators can automatically check taxonomic lists for any names that are not valid.

Ideally a nomenclator provides dereferenceable links to the treatment (part of article in which the organism was described), associated illustrations, and type materials. GNUB, the architecture for ZooBank, is designed for the role of node for nomenclators.

Annotation

Annotation can be used to fill data gaps, correct errors, and reward contributors by providing metrics for contributions, re-use, and enhancements. FilteredPush or Kurator could be adapted to this role. Annotation architecture can be linked to UUID5 identifiers minted by relevant nodes for name-strings, reconciliation groups, and other key elements. Annotation provides a pathway to grade sources as to their trustworthiness, credibility and responsiveness.

Literature subnode

A one stop shop to literary content. To include a component like RefBank that compiles citations, and normalizes them using citation parsing algorithms such as AnyStyle.io <https://anystyle.io/>. The links from citations to be dereferenceable to images / pdfs etc. of articles. This component would continue the task of digitizing and indexing biodiversity papers and books. Biodiversity Heritage Library has covered about 10% of content, and could become a node that compiles comparable content from other compilations such as BioStor or JSTOR and permit upload of user-scanned content. With enhanced biodiversity Literature Mining to extract names, collect associated data (such as location, date, and names of associated taxa), extract treatments and taxonomically relevant illustrations.

Hosting

A service oriented infrastructure must be persistent. A University, Museum, Library or similar enterprise may take responsibility to host nodes. Universities engage students in biodiversity bioinformatics. The long term host environment may be mirrored (as with Genbank and INDSC). A core team at the host is needed to provide ongoing support in software development, hardware maintenance, data management and discipline expertise. The team must harbor cutting edge understanding of how to link distributed resources, requires understanding of

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-04 15:10:16

PAGE 5

REFERENCE NO: 216

discipline standards, development of schemas to allow reasoning across proximate areas. May require involvement of TDWG and innovative approaches such as Dat (<https://datproject.org/>).

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Management and Oversight

We identify three subdomains answering to a Director, who in turn is answerable to OCI and an advisory panel .

1. Research Arm to promote discovery and invention using the conventional research paradigm to activate small groups and short term projects
2. Infrastructure arm. The production of persistent infrastructure is unfamiliar territory for an agency that is focussed on discovery and will need independent management. This needs to engage a rich mixture of stakeholders and interested parties, under a harsh agenda to prioritize new alliances and strategies. It should be well funded to issue contracts, and be tasked to emerge with budget strategies capable of sustaining an infrastructure. It will need to engage with institution/s to provide the hosting environment
3. Teaching and Training arm tasked to identify national needs and opportunities, create a training program that embraces computing science, data management, and biodiversity skills, have a budget to build portable biodiversity informatics programs

Advisory Group

Provides strategic guidance to the Director and reviews feedback on progress so that the enterprise can adopt an 'agile' approach to progress. It will represent potential components, successful enterprises, communities of sources and beneficiaries, and overseas partners. It will work with other initiatives, to urgent research targets, identify nodes and map relationships, advise on training needs and the educational agenda, promote partnerships and co-ordination nationally and internationally, set standards, guide interoperability architecture, set milestones, and promote diversity engagement

Rewarding Participation

The proposed area has the potential to become a growing infrastructure that will speed up research and add new approaches to discovery. Players be rewarded for participation (3, 4, 11, 13). Metrics of contributions made and the re-use of content may stimulate participation as will quick funding to deliver tools and services that will accelerate research agendas.

Funding Paradigm

The paradigm for research funding is not suited to the development of cyberinfrastructure. Research investment targets discovery and promotes innovation by supporting many small groups for short terms. This fosters a competitive attitude which may accelerate innovation. It is not suited to infrastructure which must focus on service and co-operation to build a persistent, valuable, common resource. The virtual nature of infrastructure also creates opportunities in communities under-represented in NSF funding because of the absence of laboratories or similar research-oriented infrastructure.

Readiness

The opportunity for a names based infrastructure has made possible because of progress in existing digital biodiversity projects and initiatives funded by NSF, such as Global Names, OTOL, ZooBank, iDigBio, BHL, EOL, and similar initiatives overseas. They can be supported to interconnect and expand into an infrastructure that can traverse about half of the envelope of biological activities. Initial implementation of the first 'proof of concept' and 'prototype' elements of a cyberinfrastructure should address urgent research challenges. They include the seamless melding of systems that rely on molecular identifiers with traditional approaches, and automated analysis of historical occurrence data to better explore origins and spread of diseases, invasive species or responses of many species to changing temperature or rainfall.

Urgency

The taxonomic community frequently identifies collections that are at risk. While writing this, the University of Louisiana announced the likely and very rapid closure of the Monroe Museum of Natural History and that 6.5 million plant and fish specimens will have to be

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-04 15:10:16

PAGE 6

REFERENCE NO: 216

donated or destroyed

http://gizmodo.com/university-threatens-destruction-of-millions-of-specime-1793745389?utm_campaign=socialflow_gizmodo_facebook&utm_source=gizmodo_facebook&utm_medium=socialflow>. An appropriate consortium making up a node can protect the scientific value in such material.

References

1. National Academy of Sciences 2009. A New Biology for the 21st Century. ISBN-13: 978-0-309-14488-9.
2. Robinson, N. 2016. Did You Know? 5 Things About BIOSIS
<http://clarivate.com/did-you-know-5-things-about-biosis/?category=science-research-connect>
3. Thessen, A. E., and Patterson, D. J. 2011. Data issues in the life sciences. ZooKeys 150: 15–51. doi: 10.3897/zookeys.150.1766
4. Patterson, D. J., Cooper, J., Kirk, P. M. and Remsen D. P. 2010. Names are key to the big new biology. TREE doi:10.1016/j.tree.2010.09.004
5. Patterson, D. J., Mozzherin, D., Shorthouse, D. P. and Thessen, A. 2016. Challenges with using names to link digital biodiversity information. Biodiversity Data Journal, doi: 10.3897/BDJ.4.e8080.
6. IUBS 2015 – Frontiers in Unified Biology <http://www.iubs.org/pdf/Events/Liste%20ReferentInnen.pdf>
7. Boyle B., et al. 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics 14 (1): 16. DOI: 10.1186/1471-2105-14-16
8. Hinchliff C.E. et al. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci U S A. 112:12764-9. doi: 10.1073/pnas.1423041112.
9. Morris, R. et al. 2013. Semantic Annotation of Mutable Data. PLoS ONE 8 (11): e76093. doi: 10.1371/journal.pone.0076093
10. Dou, L. et al., . 2012. Kurator: A Kepler package for data curation workflows. Procedia Computer Science 9, 2012, 1614-1619 <https://doi.org/10.1016/j.procs.2012.04.177>
11. Thessen, A.E., et al., 2016. Case Statement/Charter for the establishment of an Joint RDA/TDWG Working Group on Metadata Standards for attribution of physical and digital collections stewardship.
<https://www.rd-alliance.org/group/metadata-standards-attribution-physical-and-digital-collections-stewardship/case-statement>
12. Wren JD and Bateman, A. 2008. Databases, data tombs and dust in the wind Bioinformatics 24 (19): 2127-2128. DOI <https://doi.org/10.1093/bioinformatics/btn464>
13. Egloff, W., Agosti, D., Kishor, P, Patterson, D.J., Miller, J. 2017 Copyright and the Use of Images as Biodiversity Data. Research Ideas and Outcomes, 3: e12502 (06 Mar 2017) doi: 10.3897/rio.3.e12502.
14. Jones, R. 2017 (28th March). University threatens destruction of millions of specimens if Museum of Natural History collection not relocated. <http://gizmodo.com/university-threatens-destruction-of-millions-of-specime-1793745389>

Signatories

- Global Names: Dmitry Mozzherin (dmozzherin@gmail.com, University of Illinois), David Patterson (djpatterson.usyd@gmail.com University of Sydney)
- Nomenclators / ZooBank / GNUB (Rich Pyle, Bishop Museum, deepreef@bishopmuseum.org)
- Catalogue of Life in US (David Eades and Ed deWalt, University of Illinois dceades@illinois.edu, dewalt@illinois.edu)
- iDigBio (Greg Riccardi greg.riccardi@cci.fsu.edu, Larry Page lpage1@ufl.edu)
- Biodiversity Heritage Library (Martin Kalvatovic kalfatovicm@si.edu)
- World Flora Online Technical Working Group (Chair, Chuck Miller chuck.miller@mobot.org)
- User Interface (GNITE David Patterson djmapleferryman@gmail.com and TaxonWorks Matt Yoder diapriid@gmail.com)
- Concepts (Nico Franz nico.franz@asu.edu and Bertram Ludäscher ludasch@gmail.com)
- Text mining (Hong Cui hong1.cui@gmail.com)
- Map of Life (Rob Guralnick, robgur@gmail.com)
- Open Tree of Life (Jonathan Rees rees@mumble.net)
- Filtered Push / Kurator / Data Worthiness (Paul Morris mole@morris.net) James Macklin james.macklin@gmail.com, Bob Morris morris.bob@gmail.com ,Nico Franz nico.franz@asu.edu
- GenBank Detlef Leipe (leipe@ncbi.nlm.nih.gov)

Overseas

- Museum of Nature, Ottawa - Names-based services (David Shorthouse davidpshorthouse@gmail.com)
- OBKMS (RefBank, TreatmentBank, Pensoft (Donat Agosti agosti@plazi.org, Plazi, Lyubo Penev lyubo.penev@gmail.com (Pensoft Publishers)
- GBIF (Donald Hobern dhobern@gbif.org and Rod Page roderic.page@glasgow.ac.uk)

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-04 15:10:16

PAGE 7

REFERENCE NO: 216

- IUBS Hiroyuki Takeda (htakeda@bs.s.u-tokyo.ac.jp)
- TDWG Biodiversity information Standards Dimitris Koureas (d.koureas@nhm.ac.uk) John Wieckzorek (tuco@berkeley.edu)
- Institution(s) (as hosting environment or for education and training)
- Johns Hopkins University (Sayeed Choudhury sayeed@jhu.edu)
- University of Illinois (Allen Renear renear@illinois.edu)
- Florida Museum of Natural History. Map of Life (Rob Guralnick, robgur@gmail.com) and iDigBio Larry Page lpage1@ufl.edu)

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."
-